DOCUMENT RESUME

ED 421 532 TM 028 858

AUTHOR Bay, Luz
TITLE Comparing Student Performance on Different Item Formats Relative to Achievement Levels Cutpoints.
SPONS AGENCY National Assessment Governing Board, Washington, DC.
PUB DATE 1998-04-00
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Diego, CA, April 14-16, 1998).
CONTRACT ZA9003001
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Academic Achievement; Comparative Analysis; *Constructed Response; *Cutting Scores; Elementary Secondary Education; Item Response Theory; *Multiple Choice Tests; National Competency Tests; Performance Factors; Scaling; Test Format; *Test Items; Test Results
IDENTIFIERS Plausibility (Tests); Test Characteristic Curve

ABSTRACT

A study was conducted to investigate the difference in student performance on multiple choice (MC) and constructed response (CR) items relative to the achievement levels of the National Assessment of Educational Progress (NAEP). The study included an investigation of how estimates of student performance were affected by item response theory (IRT) scaling and plausible values methodology. Cutpoints were computed by panelists in the achievement levels setting process. For each grade level, seven blocks of items were selected for the study. Raw score data were provided by the Educational Testing Service for blocks from four selected test forms. The numbers of students scoring at or above each cutpoint for the respective item types and for the combination of the two item types were determined for each form. Panelists' cut point ratings were converted to the percent correct metric and the aggregate was averaged across panelists, and each cutpoint was also mapped to the percent correct metric using test characteristic curves. By either method, students performed better on MC items relative to MC cut points than on CR items relative to CR cut points. Another look at the analyses shows that for MC items, performance expectations were low relative to actual performance, while for CR items expectations were high relative to actual performance. (Contains seven tables, three figures, and five references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 421 532

Comparing Student Performance on Different Item Formats Relative to Achievement Levels Cutpoints[1]

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Luz Bay

ACT, Inc.

Luz Bay

TM028858

[1]Paper presented in M.L. Bourque (Moderator), *Setting Standards for the NAEP*, Related-Papers Session at the meeting of the National Council in Measurement in Education, April 14-16, 1998, San Diego, CA.

The research reported here was supported by contract #ZA9003001 from the National Assessment Governing Board. The study was suggested by the 1996 NAEP Science Achievement Levels-Setting project Technical Advisory Committee on Standard Setting (TACSS).

Comparing Student Performance on Different Item Formats Relative to Achievement Levels Cutpoints

Introduction

In the NAEP achievement levels-setting processes conducted by ACT the cutpoints obtained from polytomous items have generally been found to be higher than those from dichotomous items (ACT, 1993; ACT, 1995a; ACT, 1995b). In the 1996 NAEP in Science the overall cutpoints were closer to the polytomous cutpoints than to the dichotomous cutpoints (ACT, 1997). These differences in cutpoints for different item formats might be due to differences in performance, differences in the methods used to set the cutpoints, or they might just be artifacts of the "givens" in the NAEP environment. Moreover, it is possible that dichotomous items, which are almost all multiple choice (MC) items, and polytomous items, which are all constructed response (CR) items, measure different skills and knowledge. Traub and Fisher (1977) indicated that this occurrence does not generalize across subject areas, however. That is, whether tests with identical content but different formats measure the same attribute depends on the subject matter. In Traub and Fisher's (1977, p. 363) study, results "indicate that the tests of mathematical reasoning measured the same attribute regardless of response format, whereas the attributes measured by tests of verbal comprehension varied as a function of response format."

Panelist responses to ALS process evaluations indicated that 75% of the panelists agreed that CR items assess dimensions of knowledge and skills that are significantly different from those assessed by MC items (ACT, 1997). They also indicated, although not very strongly, that if ratings of student performance on MC items and CR items were very different, it was most likely caused by different student behavior and performance on the items. In a separate question, they also *somewhat agreed* that the difference in ratings might be due to the different rating methods.

The purpose of this study was to investigate the difference in student performance on MC and CR items relative to the achievement levels. The study included an investigation of how estimates of student performance were affected by Item Response Theory (IRT) scaling and plausible values methodology.

Computation of Cutpoints

During the item-rating process, each panelist estimated the expected performance on each item for students who would just meet the minimum criteria for performance at each achievement level. That is, for each multiple-choice (MC) or dichotomous item, each panelist estimated the percent of students performing at the borderline of each achievement level who would respond to the item correctly. For each polytomously scored constructed response (CR) item, each panelist estimated the average score of students performing at the borderline of each achievement level. The ratings for each item were then averaged across panelists. The average ratings were summed for each group of items based on item type and content area or subscale. Using the test characteristics curve (TCC) for each item type for each subscale, the sums of the average ratings were mapped to the theta (θ) scale. The dichotomous and polytomous cutpoints for each achievement level were then averaged to form the cutpoints for each subscale. This average was weighted, based on the amount of information at the θ value where the dichotomous and polytomous cutpoints were set. Then, cutpoints for the three fields of science were averaged and framework weights were applied. That is, if θ_{dxj} and θ_{pxj} were the dichotomous and polytomous cutpoints for achievement level x for subscale j, respectively, and if $i_{\theta dxj}$ and $i_{\theta pxj}$ were the information at the respective locations of the cutpoints, then the cutpoint for achievement level x and subscale j is given in Equation 1.

$$\theta_{xj} = \frac{i_{\theta_{dxj}}\,\theta_{dxj} + i_{\theta_{pxj}}\,\theta_{pxj}}{i_{\theta_{dxj}} + i_{\theta_{pxj}}} \qquad (1)$$

Equation 2 is the cutpoint for achievement level x, where n is the number of subscales and w_j is the framework weight for subscale j.

$$\theta_x = \frac{\sum_{j=1}^{n} w_j\,\theta_{xj}}{\sum_{j=1}^{n} w_j} \qquad (2)$$

Based on the three cutpoints (i.e., one for each achievement level) and the distribution of plausible values, the percent of students performing at or above each achievement level is estimated.

The 1996 NAEP Science ALS study was held in Phoenix, AZ in September, 1996 (ACT, 1997). The cutpoints resulting from that process are presented in Table 1. All the cutpoints were on the ACT NAEP-like scale. Although the ALS cutpoints were based on the weighted averages of the polytomous and dichotomous cutpoints, the cutpoints computed separately according to item format are presented here for MC and CR items. The rationale for was that if the differences in student performance were due to student behavior, the source of difference in behavior should be something observable to the student. Item formats (i.e., MC and CR) were observable, whether the items were scored polytomously or dichotomously was not.

In Table 1, it is very clear that cutpoints based on CR items were always higher than those based on MC items. Additionally, the overall cutpoints were always closer to the CR cutpoints. Since the performance relative to the achievement levels was substantially different for the

different item formats, the method by which cutpoints were combined might underestimate the performance of students on the NAEP.

The percentage of students scoring at or above each achievement level in Table 1 was based on the posterior distribution of student performance on both MC and CR response items and background variables. Thus, even though the cutpoints were based on MC and CR items separately, the distributions of student performance combines performance on both MC and CR items. Furthermore, items were calibrated together within subscales. This implies that the item characteristics from the estimation model were affected by the performance of the students on the combination of item types.

Data

Since the purposes of this study involved comparisons of student performance on different item types and based on different estimation protocols—not estimating student performance, per se—it was not deemed necessary to use all the 37 test forms nor to use the whole item pool for each grade level.

For each grade level, seven blocks of items were selected for this study. The collection of items in the blocks were judged to be fairly representative of each grade level item pool in terms of the distribution of items across fields of science, distribution of items across item types, and the average of the overall p-values. The seven blocks of items constituted four different test forms.[2] Information about the forms used for the study is in Table 2.

The Educational Testing Service (ETS) provided raw score data for this study. For each of the four selected forms, the number of students at each score level based on the number of

[2] There were three different types of item blocks in the 1996 NAEP Science assessment; hands-on, theme-based, and concept/problem solving blocks. Each test form was composed of three blocks. The last block was always a hands-on block. The first two blocks were either two concept/problem solving blocks, or one theme-based block and one concept/problem solving block.

points was obtained. A correct response to a MC was scored 1. For short constructed response (SCR) items, a complete response was scored 2, and a partial response was scored 1. No points were scored for incorrect, omits, not reached, and off-task responses. A response to an extended constructed response (ECR) item was scored 3, 2, 1, or 0 points. Three separate frequency distributions were used: (1) scores on MC items only; (2) scores on CR items only; and (3) scores on all items combined. Distributions of student performance, based on raw scores, relative to each achievement level cutpoint were examined. That is, the percentages of student scoring at or above each achievement level based on MC items only, and the percentage of students scoring at or above each achievement level based on CR items only were estimated using the four selected forms for each grade level.

Analyses

Because the maximum number of possible points for each form was different, the raw scores were converted to a common metric; i.e., the percent correct metric. If M were the maximum possible points for MC items only and C were the maximum possible points for CR items only, then a score of m on MC items only was converted to $100(m/M)$ and a score of c on CR items only was converted to $100(c/C)$ for that item. The total score on the percent correct metric would then be $100[(m+c)/(M+C)]$.

To examine the score distribution relative to each cutpoint, the cutpoints would have to be in the same metric. One strategy was to convert panelists' ratings to the percent correct metric and average the aggregate across panelists. These cutpoints based on raw ratings were totally free of IRT modeling. Another strategy was to map each cutpoint to the percent correct metric using test characteristic curves (TCCs). These cutpoints were, of course, influenced by IRT modeling. The percent-correct cutpoints could be computed for all items for the grade level or

only the seven blocks of items comprising the four forms used in this study. The percent correct cutpoints[3] are presented in Tables 3 and 4. The mapping of cutpoints to the percent correct metric for each grade using the TCC for all items is presented in Figures 1-3. The corresponding cutpoints were not very different, whether they were based on all items or just the seven blocks of items. This was an indication of how well the seven blocks of items represented the grade level item pool. Notice that the MC and CR cutpoints were farther apart at the Basic than at the Proficient level, and that they were closest together at the Advanced level. Finally, notice that the overall cutpoints, based on raw ratings, were consistently higher than those based on IRT estimates. Thus, the scale score predicted by item rating (i.e., percent-correct estimates from panelists) would, in turn, predict lower percentages correct than the average estimates by panelists.

Results and Discussion

The numbers of students scoring at or above each cutpoint for the respective item types and for the combination of the two item types were determined for each form. The numbers were added across forms, and the sum was divided by the total number of students who took the four forms. The percentages of students scoring at or above the cutpoints are presented in Tables 5 and 6. Because the percent correct cutpoints based on all items and the percent correct cutpoints based on the selected blocks were very similar, only the percentages of students scoring at or above cutpoints based on all items were presented.

The results presented in Table 5 are considered *free of scaling and conditioning*. Since the cutpoints were represented by raw ratings and the performances were based on raw scores,

---

[3]The percent correct cutpoint at an achievement level is interpreted as the expected percent correct score at the lower borderline of that achievement level.

then student performance relative to the cutpoints were free of IRT and plausible values methodology. The results presented in Table 6, however, were only *free of plausible values methodology*. Because the cutpoints were represented by percent-correct scores obtained using TCCs, the cutpoints were not IRT-free. The student performances considered were also based on raw scores.

For each grade, the difference between the MC and CR cutscores (i.e., percent-correct scores) consistently decreased as the level of performance increased, going from Basic to Proficient, to Advanced. This was true whether the cutpoints were based on average ratings or TCCs. On the other hand, the ratio of the percentage of students scoring above the MC cutpoints to the percentage of students scoring above the CR cutpoints increased as the level of performance increased.

In both tables, there were very strong indications that students performed much better on MC items relative to the MC cutpoints than on CR items relative to the CR cutpoints. Although the cutscore for CR items was always a lower percentage than for MC items, the performance of students relative to that cutscore was lower as well. This might have been due to performance, *per se* or to students' test taking behavior. MC items were clearly more subject to risk-taking behavior (i.e., guessing) than CR items. The effort required to respond to a CR item was generally greater than required for MC items.

Another way of looking at the results, however, is that the ratings provided by panelists on MC items, hence the MC cutpoints, were low relative to student performance on the MC items, and that the ratings provided by panelists on CR items, hence the CR cutpoints, were high relative to the student performance on CR items. In short, with MC items, performance expectations were low relative to actual performance; with CR items, performance expectations were high

relative to actual performance. This would be the case even though the CR cutpoints were always lower than the MC cutpoints.

For two of the four forms used for grades 4 and 8, plausible value scores for all students who took the forms were available for comparing student performance relative to achievement levels based on conditioned scores and raw scores. The percentages of students scoring at or above each achievement level based on raw scores and plausible values were reported in Table 7.

In grade 4, performance seems to have increased with each additional psychometric application. That is, for the Basic and Proficient achievement levels, student performance relative to average raw score ratings was lowest, and student performance based on plausible values relative to actual cutpoints was highest. In grade 8, this was true only at the Basic level.

The large difference between student performance relative to the expected score cutpoints and student performance based on the actual cutpoints seemed to indicate a substantial effect of conditioning on student performance in grade 4. Such was not the case in grade 8. This finding is somewhat ironic, in that the **student-reported** background data used in conditioning are generally regarded to be less reliable for grade 4 compared to other grades.

Reference

ACT (1997). *Setting achievement levels on the 1996 National Assessment of Educational Progress in Science: Final report.* Iowa City, IA: Author.

ACT (1995a). *Results of the 1994 Geography NAEP Achievement Levels-Setting pilot study.* Iowa City, IA: Author

ACT (1995b). *Results of the 1994 U.S. History NAEP Achievement Levels-Setting pilot study.* Iowa City, IA: Author

ACT (1993). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing: A technical report on reliability and validity.* Iowa City, IA: Author.

Traub, R.E. and Fisher, C.W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1,* 355-369.

Table 1

Numerical Results of the 1996 NAEP Science ALS Process

Grade	Item Type	Achievement Levels					
		Basic		Proficient		Advanced	
		Cutpoint (SD)	% ≥	Cutpoint (SD)	% ≥	Cutpoint (SD)	% ≥
4	Multiple-Choice	138.0 (9.5)	88.3	161.9 (6.7)	30.9	181.2 (6.0)	0.9
	Constructed Response	144.0 (4.5)	79.4	168.3 (2.9)	14.6	189.4 (4.7)	0.1
	Combined	**142.6 (4.4)**	**82.2**	**166.9 (2.6)**	**17.3**	**187.4 (3.7)**	**0.1**
8	Multiple-Choice	145.7 (15.6)	75.1	170.2 (9.9)	11.6	189.3 (9.7)	0.1
	Constructed Response	156.3 (9.9)	48.3	178.1 (8.7)	2.8	196.6 (8.9)	0.0
	Combined	**154.2 (10.1)**	**54.2**	**176.7 (7.9)**	**3.9**	**195.5 (8.4)**	**0.0**
12	Multiple-Choice	151.0 (9.0)	62.4	167.6 (4.4)	16.5	182.0 (4.3)	1.2
	Constructed Response	156.3 (6.1)	47.8	175.0 (4.7)	5.2	193.6 (5.3)	0.0
	Combined	**154.6 (6.0)**	**52.4**	**173.0 (2.9)**	**7.5**	**188.3 (3.5)**	**0.3**

Bold indicates information presented to the panelists.

Table 2

Information on Items Used for the Study .

Blocks	Number of Blocks				Number of Items (%)						Average P-Value		
	All	Block Type			All	Field of Science			Item Format		All Items	MC Items Only	CR Items Only
		Hands-On	Theme-Based	Concept/ Problem Solving		Physical	Earth	Life	Multiple-Choice	Constructed Response			
Grade 4													
All	15	4	3	8	136	43 (32)	49 (36)	44 (32)	50 (37)	86 (63)	45	57	38
Study	7	2	1	4	64	21 (33)	21 (33)	22 (34)	25 (39)	39 (61)	44	58	35
Form 1*	3	1	1	1	24	7 (29)	6 (25)	11 (46)	6 (25)	18 (75)	37	60	29
Form 2	3	1	1	1	26	10 (38)	4 (15)	12 (46)	8 (31)	18 (69)	44	58	38
Form 3	3	1	1	1	26	5 (19)	10 (38)	11 (42)	9 (35)	17 (65)	45	67	34
Form 4*	3	1	1	1	26	9 (35)	5 (19)	12 (46)	8 (31)	18 (69)	44	46	43
Four Forms Combined**					102	31 (30)	25 (24)	46 (45)	31 (30)	71 (70)	43	58	36
Grade 8													
All	15	4	3	8	190	62 (33)	64 (34)	64 (34)	72 (38)	118 (62)	41	55	33
Study	7	2	1	4	88	30 (34)	28 (32)	30 (34)	38 (43)	50 (57)	41	56	30
Form 1*	3	1	1	1	32	8 (25)	6 (19)	18 (56)	13 (41)	19 (59)	39	60	25
Form 2	3	1	0	2	40	16 (40)	16 (40)	8 (20)	17 (43)	23 (58)	43	59	31
Form 3	3	1	0	2	41	17 (41)	16 (39)	8 (20)	17 (41)	24 (59)	44	55	36
Form 4*	3	1	0	2	37	15 (41)	12 (32)	10 (27)	16 (43)	21 (57)	38	48	31
Four Forms Combined**					150	56 (37)	50 (33)	44 (29)	63 (42)	87 (58)	41	55	31
Grade 12													
All	15	4	3	8	186	59 (32)	62 (33)	65 (35)	69 (37)	117 (63)	45	57	38
Study	7	2	1	4	89	33(37))	26 (29)	30 (34)	35 (39)	54 (61)	43	53	36
Form 1	3	1	1	1	39	12 (31)	16 (41)	11 (28)	13 (33)	26 (67)	43	58	35
Form 2	3	1	1	1	35	12 (34)	16 (46)	7 (20)	13 (37)	22 (63)	47	56	42
Form 3	3	1	0	2	40	16 (40)	6 (15)	18 (45)	16 (40)	24 (60)	42	47	38
Form 4	3	1	0	2	34	15 (44)	7 (21)	12 (35)	14 (41)	20 (59)	41	52	34
Four Forms Combined**					148	55 (37)	45 (30)	48 (32)	56 (37)	92 (62)	43	53	37

s presented in Table 7.

* Forms used for result

** The number of items is based on the number of times they are used in the four forms.

Table 3

Percent Correct Cutpoints Based on Raw Ratings

Grade	Items	Format	Basic	Proficient	Advanced
4	All	MC	40.0	65.2	84.1
		CR	24.5	52.3	75.4
		Combined	30.8	57.5	78.9
	Selected Blocks	MC	39.6	65.7	84.6
		CR	25.4	52.6	74.8
		Combined	31.1	57.7	78.7
8	All	MC	47.6	71.2	86.7
		CR	29.4	58.4	79.7
		Combined	36.9	63.7	82.6
	Selected Blocks	MC	48.4	71.5	87.0
		CR	27.6	56.3	77.8
		Combined	37.6	63.6	82.2
12	All	MC	51.7	73.4	89.1
		CR	35.0	62.7	83.5
		Combined	41.9	67.2	85.8
	Selected Blocks	MC	49.5	72.0	88.6
		CR	35.6	63.6	84.0
		Combined	41.7	67.2	86.0

Table 4

Percent Correct Cutpoints Based on Test Characteristic Curves

Grade	Items	Format	Basic	Proficient	Advanced
4	All	MC	40.6	64.0	82.7
		CR	24.2	51.9	75.1
		Combined	28.2	54.8	76.4
	Selected Blocks	MC	39.3	62.6	81.0
		CR	24.8	51.5	73.9
		Combined	27.8	52.5	74.2
8	All	MC	45.4	69.7	86.6
		CR	30.2	60.0	80.9
		Combined	34.6	62.8	82.2
	Selected Blocks	MC	47.4	70.6	87.4
		CR	26.8	58.1	79.8
		Combined	29.9	54.0	76.2
12	All	MC	49.6	71.9	88.8
		CR	33.9	63.1	83.4
		Combined	37.6	64.5	82.1
	Selected Blocks	MC	47.7	73.2	89.8
		CR	34.9	64.2	84.8
		Combined	36.6	64.2	81.2

Table 5

Estimated Percentages of Students Scoring At or Above Each Achievement Level Based on Average Ratings

Grade	Items	Basic		Proficient		Advanced	
		Average Ratings	%≥	Average Ratings	%≥	Average Ratings	%≥
4	MC Only	40.0	85.05%	65.2	51.12%	84.1	16.03%
	CR Only	24.5	73.31	52.3	12.51	75.4	0.18
	All	30.8	67.92	57.5	9.88	78.9	0.18
8	MC Only	47.6	74.10	71.2	25.69	86.7	6.25
	CR Only	29.2	49.41	58.4	4.4	79.7	0.00
	All	36.9	48.96	63.7	16.58	82.6	0.17
12	MC Only	51.7	56.80	73.4	18.65	89.1	5.61
	CR Only	35	53.12	62.7	7.41	83.5	0.21
	All	41.9	46.06	67.2	6.38	85.8	0.21

Table 6

Estimated Percentages of Students Scoring At or Above Each Achievement Level Based on Expected Percent Correct Score

Grade	Items	Basic		Proficient		Advanced	
		% Correct Score	% ≥	% Correct Score	% ≥	% Correct Score	% ≥
4	MC Only	40.6	85.05	64.00	51.12	82.7	21.5
	CR Only	24.2	76.32	51.9	12.51	75.1	0.18
	All	28.2	73.61	54.8	14.26	76.4	0.41
8	MC Only	45.4	78.37	69.7	25.69	86.6	6.25
	CR Only	30.2	49.41	60.0	3.97	80.9	0.00
	All	34.6	53.65	62.8	4.77	82.2	0.17
12	MC Only	49.6	61.78	71.9	25.23	88.8	5.61
	CR Only	33.9	53.92	63.1	7.12	83.4	0.21
	All	37.6	55.36	64.5	7.52	82.1	0.21

Table 7[4]

Percent of Students Scoring At or Above Each Achievement Level Based on Raw Ratings, Expected Percent Score, and Plausible Values for Selected Test Forms

Grade	Items	Basic			Proficient			Advanced		
		Average Ratings	Expected Score	Plausible Values	Average Ratings	Expected Score	Plausible Values	Average Ratings	Expected Score	Plausible Values
4	MC Only	84.9	84.9		58.2	58.8		21.6	32.4	
	CR Only	59.5	65.5		5.9	5.9		0	0	
	All	58.8	66.0	78.71	6.2	8.4	16.82	0	0.4	0.21
8	MC Only	70.3	79.0		25.9	25.6		6.2	6.2	
	CR Only	45.2	45.2		3.2	2.8		0	0	
	All	45.0	50.3	52.17	3.7	3.7	2.98	0	0	0

[4] Data reported under "Average Ratings" were percentages at or above each achievement level that were similar to the ones in Table 5. Data reported under "Expected Score" were the percentages at or above similar to those in Table 6. Data reported under "Plausible Values" are the percentages of plausible value score that are above each cutpoint on the ACT NAEP-Like scale. Since there were virtually no observations at the Advanced level, comparisons of results were made only at the Basic and Proficient levels.

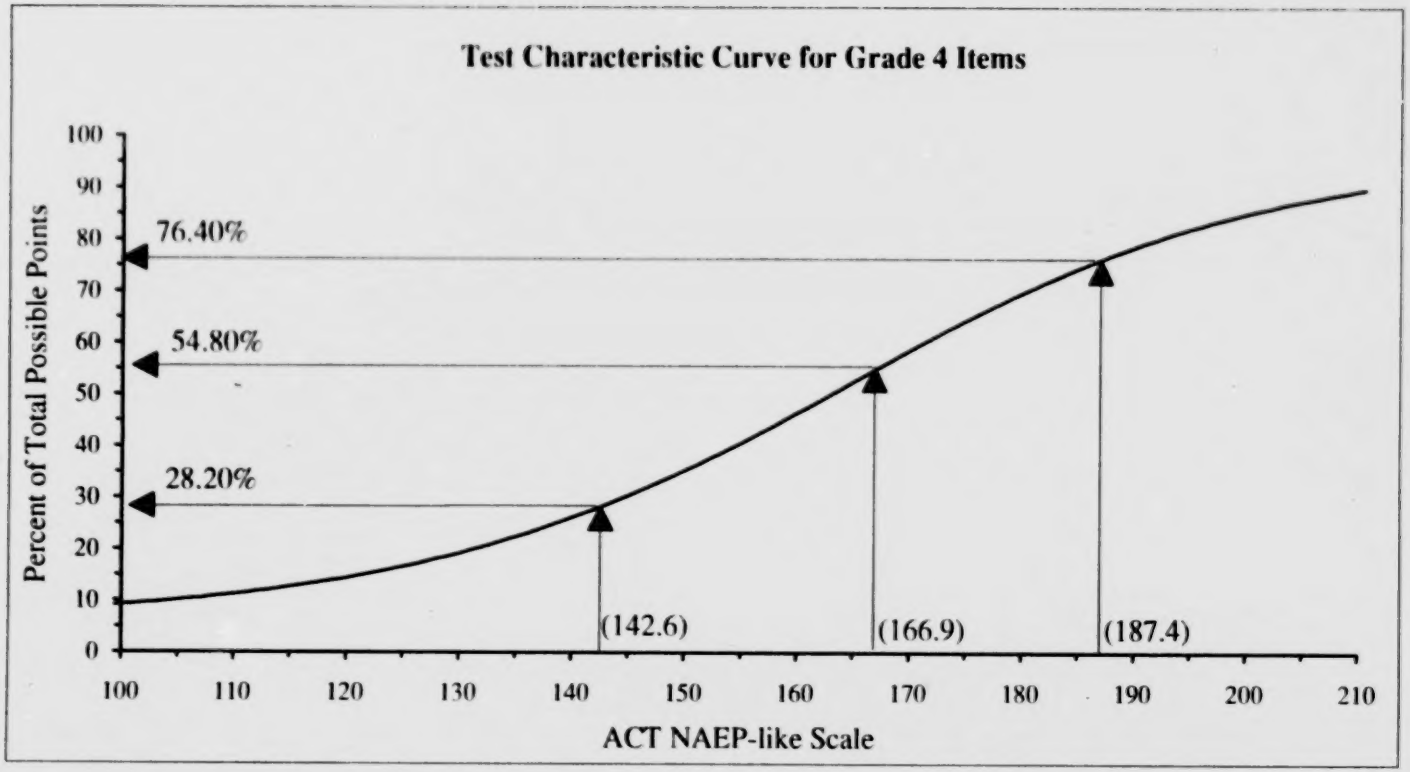

Figure 1: Expected percent-correct score at each achievement level cutpoint was estimated using the test characteristic curve for grade 4 items.

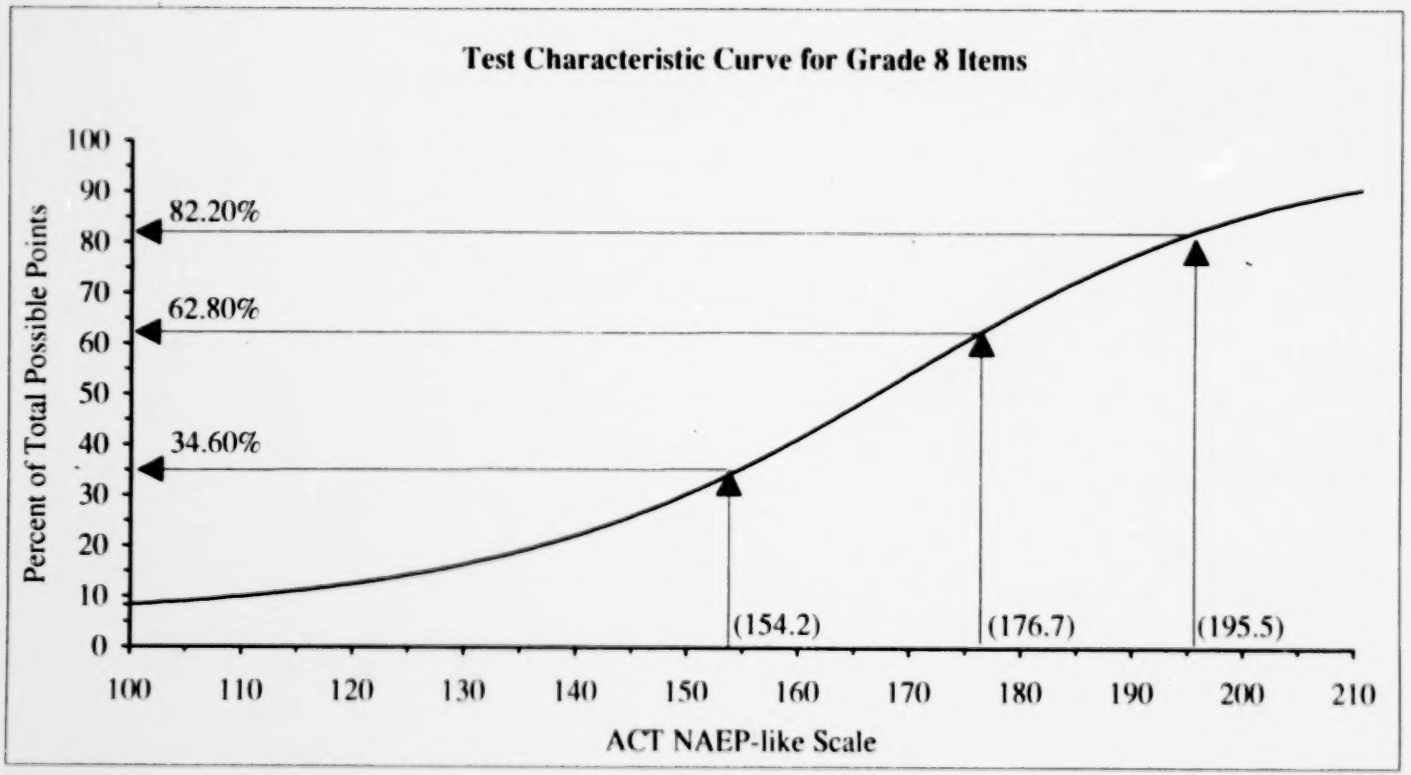

Figure 2: Expected percent-correct score at each achievement level cutpoint was estimated using the test characteristic curve for grade 8 items.

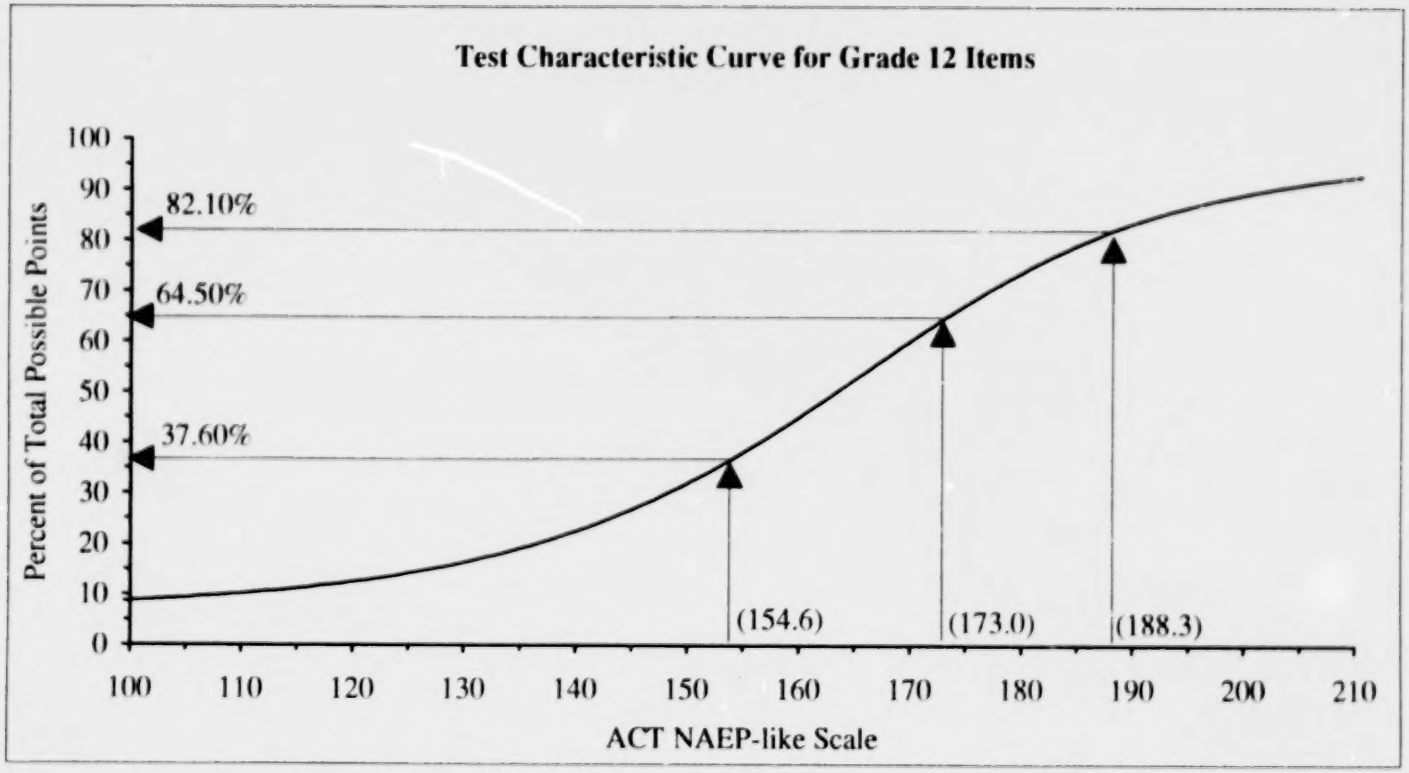

Figure 3: Expected percent-correct score at each achievement level cutpoint was estimated using the test characteristic curve for grade 12 items.